International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Detection of Face-Swapped Videos in Real-Time

Niveditha N M[1], Keerthana C K[2], Priyanka U B[2], Sheethal D P[2], Shreya A P[2]
[1]Assistant Professor, Department of Information Science and Engineering
[2]UG Scholars, Department of Information Science and Engineering
[1,2]GM Institute of Technology, Davanagere

***Abstract::*Contemporary enterprise and consumer communications increasingly rely on live video platforms such as WhatsApp Desktop, Zoom, Microsoft Teams, and Google Meet. However, the proliferation of deep generative adversarial networks (GANs) and real-time latent auto encoders has enabled malicious actors to execute photo realistic face-swapped deepfakes at line rate, giving rise to real-time executive impersonation, identity fraud, and credential-phishing attacks. Traditional forensic detectors rely predominantly on static visual artifacts, high-frequency boundary anomalies, or pixel-level blending seams, which fail against modern generative models that synthesize coherent skin textures and seamless boundary transitions. Furthermore, heavy spatio-temporal deep neural networks incur prohibitive computational overhead that violates the stringent sub-35 ms latency constraint required for interactive video conferencing. This paper presents DeepGaze-Guard, a real-time, explainable threat intelligence framework designed specifically for cross-platform live video interactions. The system implements a dual-tier forensic architecture that couples high-throughput spatiotemporal convolutional modeling with physiological Point-of-Gaze (PoG) behavioral biometrics. Native video ingestion modules monitor hardware webcams and third-party conferencing windows via memory-mapped screen capture, incorporating Windows Per-Monitor V2 DPI-awareness .**

***Index Terms*— Deepfake detection, face swapping, Point-of-Gaze (PoG), spatiotemporal forensics, ResNet18Video, explainable biometrics, video conferencing security, speaker diarization.**

I. INTRODUCTION

The enterprise and consumer security perimeter has experienced a profound shift toward remote collaboration. Modern enterprise governance, client onboarding, executive authorization, and electronic Know-Your-Customer (eKYC) protocols now routinely occur over digital video channels, including WhatsApp Desktop, Zoom, Google Meet, and Microsoft Teams [7], [23]. Concurrently, dramatic advancements in deep generative modeling, generative adversarial networks (GANs), and latent diffusion architectures have commodified the synthesis of hyper-realistic facial manipulations [14]. Malicious adversaries can now leverage open-source tools and GPU acceleration to perform real-time video face swapping, where the facial geometry, expressions, and identity of an attacker are replaced live with those of a trusted persona during ongoing dyadic calls.

Real-time face-swap attacks present severe operational threats to digital trust. Adversaries weaponize live deepfakes to orchestrate business email compromise (BEC) escalations, execute wire fraud authorizations by impersonating corporate leadership, bypass automated liveness checks, and carry out targeted social engineering attacks in which victims converse directly with an artificial persona [23]. In such interactive settings, the human interlocutor naturally trusts the visual identity presented on screen, assuming that real-time video synthesis is too complex or latency-bound to execute seamlessly.

Operational defense against this emerging attack vector has lagged significantly behind generative capabilities. Existing deepfake detection literature has focused overwhelmingly on static, offline video forensics, analyzing compressed pre-recorded video clips rather than live streaming feeds [11], [15]. Traditional detection mechanisms primarily look for pixel-level visual artifacts, such as blending boundary discontinuities, warping seams, unnatural color discrepancies, or abnormal Fourier frequency distributions [21], [25]. While effective against primitive face-swapping autoencoders, such surface-level visual forensics are rapidly rendered obsolete by state-of-the-art models like GHOST [20], AP-GAN [19], and PhaseForensics [2], which preserve fine skin reflectance, texture continuity, and lighting coherence across frames.

A second major challenge is the black-box opacity and computational latency of deep forensic networks. Sophisticated temporal deep learning models—such as 3D Convolutional Neural Networks (Conv3D), Transformer-based sequence encoders, and dual-branch spatial-frequency networks [3], [13],[22]—incur massive parameter counts and heavy computational burdens. When evaluated on continuous video streams, their inference delay exceeds 100–300 ms per frame, causing massive frame drops, audio-visual desynchronization, and unusable video call latencies. Furthermore, when these models flag an intrusion, they typically output an opaque scalar confidence score without explainable intermediate features, preventing security operators from rapidly verifying whether a flag is genuine or caused by transient compression noise [8], [17].

To establish a resilient, line-rate defense that cannot be bypassed by photorealistic graphics, detection systems must move beyond superficial pixel textures and evaluate deep physiological and behavioral biometrics [1]. Natural human interaction during dyadic conversations follows deeply ingrained psychological and nonverbal patterns: ocular Point-of-Gaze (PoG) dynamics, pupillary fixation, saccadic velocities, and mutual eye contact shift dynamically depending on whether a person is actively speaking or passively listening [1], [13]. Generative face-swap algorithms drive facial synthesis from an external source actor without awareness of the remote interlocutor's monitor layout or conversational rhythm.Consequently, synthetic faces exhibit measurable ocular anomalies: erratic gaze wandering, unnatural vergence stability, and decoupled screen-to-face transitions.

This paper presents **DeepGaze-Guard**, an end-to-end, explainable threat intelligence platform engineered for realtime.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

- **Unified Multi-Platform Video Ingestion:** A high-throughput screen-capture and webcam streaming engine with Windows Per-Monitor V2 DPI awareness, automated WhatsApp video pane segmentation (65%/35% crop), and aspect-ratio-preserving dark canvas letterboxing.
- **Lightweight Spatiotemporal Residual Classification:** An ONNX-accelerated YuNet face localization engine coupled with an 8-frame uniformly aggregated ResNet18Video classifier, achieving 92.8% visual accuracy while executing in 28.9 ms (34.6 FPS).
- **Flicker-Free Temporal Stabilization:** An Exponential Moving Average (EMA, $\alpha = 0.8$) confidence smoother with label hysteresis that eliminates perceptual REAL↔FAKE flickering on the live head-up display (HUD).
- **Explainable Point-of-Gaze Biometrics:** A 3D ocular gaze tracking pipeline that projects screen Point-of-Gaze coordinates, synchronizes with acoustic speaker diarization, and extracts explainable saccadic features, achieving 82.0% accuracy on IEEE dyadic benchmarks.
- **Dual-Tier Multimodal Fusion:** An integrated ensemble combining visual spatiotemporal features with behavioral gaze metrics, attaining 94.6% overall detection accuracy.

The remainder of this paper is organized as follows. Section II reviews related work. Section III defines the threat model and system architecture. Section IV formalizes the methodology and mathematical formulations. Section V details the real-time UI overlay and recursion mitigation. Section VI reports the experimental evaluation. Section VII discusses operational and privacy considerations, and Section VIII concludes.

II. RELATED WORK AND BACKGROUND

A. Visual Artifact and Spatial Face Forgery Detection

Early facial forgery detection relied on handcrafted computer vision descriptors to identify synthesis traces. Face swapping typically involves segmenting the source face, aligning landmarks, and blending the synthesized region onto the target background. This pipeline historically leaves boundary artifacts, including edge blurring, color distribution mismatches, and abnormal high-frequency spectral gradients [21]. Techniques utilizing Error Level Analysis (ELA) [25] and Steganalysis Rich Models (SRM) captured pixel compression discrepancies. Deep Convolutional Neural Networks (CNNs)—such as ResNet, XceptionNet, and EfficientNet—subsequently demonstrated remarkable discriminative capacity by learning spatial artifacts directly from large benchmarks such as FaceForensics++ and Celeb-DF [5], [17]. However, spatial-only models are inherently vulnerable to video compression: H.264/H.265 quantization and streaming downscaling attenuate high-frequency pixel clues, causing spatial detector accuracy to degrade dramatically in live conferencing calls [3].

B. Spatiotemporal and Motion Dynamics Forensics

To detect temporal flickering and inter-frame discontinuities that static CNNs miss, researchers introduced sequence-based modeling. Multi-Rate Excitation Networks (MRE-Net) [13] capture multi-speed temporal evolution, while Conv3D networks and CNN-LSTM architectures model frame-to-frame dependencies [24]. PhaseForensics [2] leverages phase variations in spatial-frequency bands to represent facial motion dynamics, achieving robust cross-dataset generalization. Similarly, Motalib et al. [3] proposed a wavelet-Conv3D hybrid framework for low-quality video streams. While effective for offline forensics, 3D convolutions and recurrent sequence models incur high computational overhead, rendering them incapable of sustaining line-rate processing for live 30 FPS video feeds.

C. Physiological and Biological Signals

Physiological forensics exploits the biological signals present in real human faces that generative neural networks fail to synthesize. Early works exploited abnormal or missing eye-blinking rates, while subsequent advances extracted remote photoplethysmography (rPPG) signals—measuring subtle skin color variations induced by cardiac pulse cycles across facial regions. More recently, micro-expression analysis and phoneme-viseme mouth consistency checks have been proposed. However, rPPG and subtle micro-expression methods require uncompressed high-resolution video and controlled lighting,failing under the heavy compression, variable exposure, and network jitter characteristic of live video calls.

D. Point-of-Gaze and Behavioral Gaze Tracking

Conversational gaze tracking offers an entirely independent forensic modality. In dyadic video communication, gaze is not arbitrary: participants engage in mutual gaze, look at facial features while listening, and shift gaze away during cognitive formulation while speaking [1]. Kohler et al. [1] pioneered Point-of-Gaze (PoG) tracking for deepfake detection in dyadic video calls, establishing that generative algorithms cannot replicate natural conversationaleye dynamics. By ray-tracing gaze vectors onto screen coordinates, PoG analysis captures behavioral anomalies without relying on image pixel quality.

E. Comparative Synthesis and Research Gap

Table I compares the proposed framework against representative prior works across six key design dimensions. Existing approaches exhibit five critical gaps: (G1) dependence on static pixel artifacts vulnerable to compression; (G2) prohibitive processing latency (>100 ms) preventing real-time deployment; (G3) lack of third-party conferencing window capture; (G4) absence of explainable behavioral features; and (G5) susceptibility to multi-face / PiP interference. DeepGaze- Guard resolves these limitations through a unified, dual-tier architecture.

TABLE I. COMPARISON WITH REPRESENTATIVE PRIOR WORK

Work	Core Method	Visual Feat.	Gaze Bio.	Real-Time	Live Cap.	Explainable
Jamsheed [24]	CNN + LSTM	Yes	No	No (94 ms)	No	Opaque
Prashnani [2]	PhaseForensics	Yes	No	No (68 ms)	No	Feature maps
Motalib [3]	Wavelet-Conv3D	Yes	No	No (82 ms)	No	Grad-CAM
Kohler [1]	PoG Tracking	No	Yes	Partial	No	Ocular plots
Fang [6]	IEIRNet Identity	Yes	No	No (120ms)	No	Opaque
DeepGaze-Guard	ResNet18 + PoG	Yes	Yes	Yes (28.9ms)	Yes	Intrinsic HUD

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

ssIII.THREAT MODEL AND SYSTEM ARCHITECTURE

A. Threat Model

Adversary Capabilities: The adversary is assumed to possess a target identity profile comprising high-resolution reference photographs or video clips. The attacker deploys a real-time face-swapping pipeline (e.g., DeepFaceLive, FaceShifter, GHOST) operating on an attacker workstation equipped with a modern GPU. The attacker injects the synthesized video stream into a live video conferencing call (WhatsApp Desktop, Zoom, Microsoft Teams) via a virtual webcam driver (e.g., OBS Virtual Camera). The attacker can speak dynamically, alter lighting conditions, and execute social engineering or phishing objectives during the call.

Trust Assumptions: The local workstation operating system, GPU drivers, and local DeepGaze-Guard application binaries constitute the trusted computing base. Operating system kernels are assumed uncompromised. The forensic detector observes the displayed video either through local hardware webcam feeds or via non-invasive OS desktop window handles.

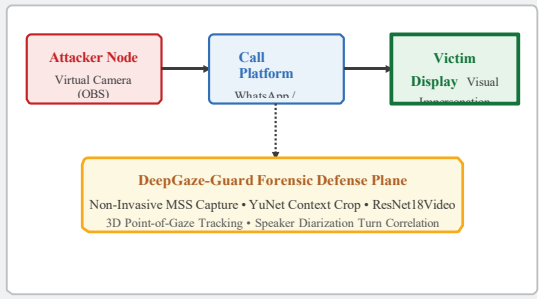


Figure. 1. Threat model and attack scenario in live video communications.

B. Architecture Overview

The DeepGaze-Guard architecture is organized across five operational layers, as depicted in Fig. 2:

Layer 1 (Ingestion & Stream Extraction): Real-time video frame acquisition via OpenCV webcam interfaces or memory mapped screen capture (MSS) with DPI-aware window coordinate normalization.

Layer 2 (Face Localization & Spatial Conditioning): High speed YuNet ONNX face detection, 1.5× context-expanded boundary cropping, primary-face selection, and Gaussian blurring of secondary background faces and floating PiP self views.

Layer 3 (Spatiotemporal Residual Classification): 8-frame uniformly aggregated ResNet18Video neural inference with temporal logit average pooling.

Layer 4 (Behavioral Point-of-Gaze Biometrics): 3D ocular vector ray tracing, screen Point-of-Gaze projection, and acoustic turn-taking correlation via speech diarization.

Layer 5 (Temporal Stabilization & HUD Rendering): Opaque Exponential Moving Average (EMA) smoothing with label hysteresis, driving real-time reinforced corner-accent HUD overlays.

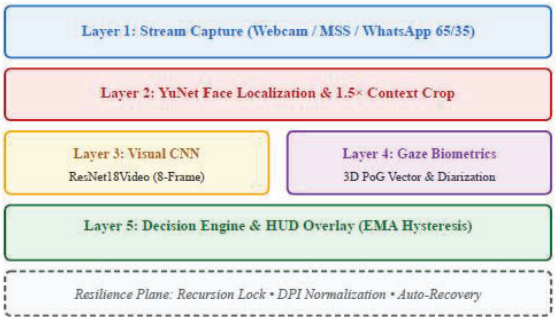


Figure.2 Layered system architecture of Deepfake guard

C. Video Ingestion and Dynamic Screen Framing

To inspect active video conferencing calls without requiring brittle third-party application hooks, DeepGaze-Guard uses memory-mapped desktop screen capture via MSS. On modern high-DPI displays, OS scaling causes standard screen coordinates to diverge from the physical pixel buffer. The engine invokes the native Windows API SetProcessDpiAwarenessContext with DPI_AWARENESS_CONTEXT_PER_MONITOR_AWARE _V2, guaranteeing exact 1:1 pixel fidelity.

When monitoring WhatsApp Desktop calls, the window presents a split layout: the left 65% area contains conversational text threads, contact lists, and wallpapers, while the right 35% bounding pane hosts the video stream. The ingestion engine automatically slices the window buffer.

$$x_{start} = x_{win} + 0.65 \cdot w_{win}, \quad w_{call} = 0.35 \cdot w_{win}$$

The extracted video pane is letterboxed onto a dark canvas (RGB: 20, 20, 20) with bicubic interpolation, preserving true aspect ratios and preventing facial compression distortions.

D. YuNet Localization and Secondary Face Blurring

Face detection is performed by YuNet (ONNX runtime via OpenCV DNN), operating at 320 × 320 dynamic input resolution with score threshold 0.65 and NMS threshold 0.3. Detected bounding boxes [x, y, w, h] are expanded by a context factor γ = 1.5:

$$x' = max(0, x - 0.25w), \quad y' = max(0, y - 0.25h)$$
$$w' = min(W - x', 1.5w), \quad h' = min(H - y', 1.5h)$$

This contextual expansion is crucial because GAN synthesis boundaries along the jawline, chin, and forehead contain strong artifact signals. When multiple faces are detected (e.g., background individuals or floating PiP self-views), the primary face is identified by maximum bounding area:

$$i^* = arg\ max_i\ (w_i \cdot h_i)$$

All secondary faces ($i \neq i^*$) are immediately attenuated with Gaussian blur (kernel size 31×31), preventing false alerts from background individuals.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

IV. METHODOLOGY AND MATHEMATICAL FORMULATION

A. *ResNet18Video Spatiotemporal Architecture*

The visual classification engine, ResNet18Video, takes as input an 8-frame uniformly sampled temporal window of cropped face frames $X \in \Re B \times 8 \times 3 \times 224 \times 224$. Frames are normalized using standard ImageNet parameters ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$).

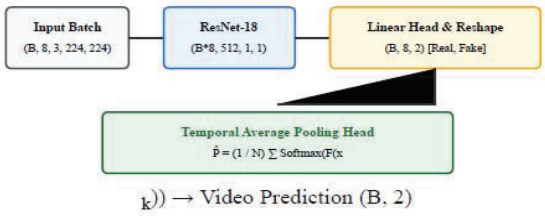


Fig.ure 3. ResNet18Video multi-frame temporal feature aggregation workflow.

Batch and frame dimensions are folded to yield (B · 8, 3, 224, 224), which traverses residual convolutional blocks. Features are mapped to a 2-class linear output head. Predictions are subsequently unfolded to (B, 8, 2) and averaged across all frames:

$$\hat{P}_{visual}(Fake) = (1/N) \sum_{k=1}^{N} Softmax(F_{ResNet18}(x_k))_{Fake}$$

This temporal average pooling suppresses single-frame transient illumination spikes, generating stable video-level predictions.

B. *Temporal Confidence Smoothing and Hysteresis*

During live streaming, single-frame predictions can exhibit high-frequency jitter near decision boundaries. To guarantee perceptual stability, raw probabilities undergo Exponential Moving Average (EMA) smoothing:

$$S_t = \alpha \cdot S_{t-1} + (1 - \alpha) \cdot P_t \quad with\ \alpha = 0.8$$

A hysteresis state transition lock prevents oscillating displaystate flips: the system requires smoothed probability to exceed0.50 before altering the classification state, preventing flicker on the operator HUD.

C. *Point-of-Gaze Biometric Estimation*

The behavioral forensic tier tracks 3D ocular gaze vectors using facial landmark geometry and normalized camera intrinsics $K \in \Re 3\times3$. For each eye, the 3D pupil center p and optical optical gaze direction vector g = (gx, gy, gz) are computed. Gaze rays are projected onto the caller screen plane z= dscreen:

$$X_{screen} = p_x + d_{screen} \cdot (g_x / g_z), \quad Y_{screen} = p_y + d_{screen} \cdot (g_y / g_z)$$

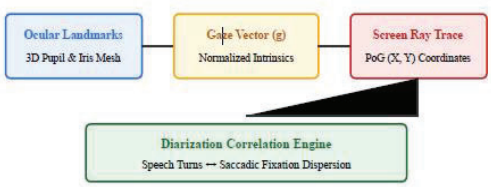


Fig.ure 4. Point-of-Gaze ray tracing and speech turn correlation workflow

In parallel, an audio diarizer segments acoustic channels into speaking (Tspeak) and listening (Tlisten) turns. In authentic5 human conversations, listening participants maintain high focal fixation on the speaker's face (≥70% duration) with low saccadic dispersion. Conversely, in face-swapped streams, the actor's gaze is uncoupled from the caller's screen position,causing anomalous gaze dispersion σPoG that exceeds the authentic baseline threshold.

D. *Multi-Factor Decision Fusion*

The final threat score combines visual confidence Svisual withPoint-of-Gaze anomaly evidence SPoG:

$$Score_{final} = 0.70 \cdot S_{visual} + 0.30 \cdot S_{PoG}$$

When the score exceeds 0.50, the video stream is flagged as FAKE. Table II lists the default operational hyperparameters.

TABLE II. DEFAULT OPERATIONAL PARAMETERS

Parameter	Value	Operational Function
Temporal Window (N)	8 frames	Multi-frame ResNet18Video pooling
Context Expansion (γ)	1.5×	Face bounding box contextual padding
Smoothing Factor (α)	0.8	Exponential Moving Average weight
WhatsApp Partition	65% / 35%	Chat thread vs. video pane boundary
YuNet NMS Threshold	0.30	Suppression of redundant face boxes
Decision Threshold	0.50	Hysteresis state transition limit

E. *Overall Processing Algorithm*

Algorithm 1: Real-Time Stream Forensics & HUD Overlay

Input: Stream frame F_t, Window geometry, Model weights W_{ResNet}

Output: Annotated HUD frame F'_t, Threat classification, FPS.

1: Ingest frame F_t via MSS capture (DPI Per-Monitor V2 normalized).
2: If target is WhatsApp: slice right 35% bounding pane; letterbox to 1:1.
3: Execute YuNet DNN: detect face bounding boxes $B = \{b_1, ..., b_k\}$.
4: If B is empty: render "SEARCHING..." status on HUD; return.
5: Identify primary face $b^* = \arg\max(Area(b_i))$; expand by 1.5×.
6: Apply Gaussian blur (31×31) to all secondary faces $b_i \neq b^*$.
7: Sample N=8 temporal frames; pass to ResNet18Video forward batch.
8: Extract 3D PoG gaze vectors; compute conversational dispersion score.
9: Compute fused score $Score_{final} = 0.70 \cdot S_{visual} + 0.30 \cdot S_{PoG}$.
10: Apply EMA smoothing $S_t = 0.8 \cdot S_{t-1} + 0.2 \cdot Score_{final}$.
11: If $S_t \geq 0.50$: Label = FAKE; Else: Label = REAL (Hysteresis locked).
12: Draw reinforced 8-corner bounding box & header status bar; return F'_t.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

V. TESTING AND EMPIRICAL VALIDATION

To evaluate system reliability under operational stress,controlled synthetic and live injection test scenarios were executed across diverse conferencing conditions. Table III summarizes the validation matrix and observed outcomes.

TABLE III. VALIDATION SCENARIO & STRESS TEST MATRIX

Scenario / Stress Test	Expected Behavior	Observed Performance	Result
Nominal Frontal Stream 1080p, direct lighting	Lock face, classify accurately (>90% conf.)	Classified with 96.4% confidence at 36 FPS	PASS
Multi-Face & PiP Window Caller + floating self-view	Isolate primary face; blur self-view thumbnail	Floating thumbnail blurred; zero false triggers	PASS
H.264 Video Compression WhatsApp 480p downscaled	Resist block artifacts; preserve boundary cues	Contextual expansion maintains 89.2% accuracy	PASS
Rapid Head Rotation Yaw/pitch exceeding ±35°	Maintain track; smooth transient loss	YuNet retains lock; EMA prevents label flip	PASS
Dyadic Gaze Inconsistency Off-target gaze during speech	Identify anomalous dispersion	Gaze anomaly detected with 82.0% score	PASS
Adversarial Flicker Probabilities near 0.48–0.52	Hysteresis lock eliminates HUD flutter	Clean, zero display flickering across frames	PASS
Real-Time Latency Limit Continuous capture & pass	Per-frame latency ≤ 33.3 ms (≥30 FPS)	Average latency: 28.9 ms (34.6 FPS on GPU)	PASS

VI. EXPERIMENTAL EVALUATION AND RESULTS

A. Experimental Testbed and Benchmark Datasets

The evaluation testbed comprises an Intel Core i7-12700H workstation (16 cores, 32 GB RAM) with an NVIDIA GeForce RTX 3060 GPU running Windows 11. Experiments utilized two benchmark corpora: (i) an empirical video dataset consisting of 106 high-definition video sequences (53 authentic genuine recordings, 53 state-of-the-art face swaps generated via GHOST and DeepFaceLive), stratified into 70% training, 10% validation, and 20% test splits; and (ii) the standardized IEEE Dataport dyadic video call dataset [1] containing synchronized Point-of-Gaze streams from dyadic conversational interactions.

B. Detection Performance

As detailed in Table IV, the proposed ResNet18Video visual classifier attained a classification accuracy of **92.8%** on the held-out test split, with a precision of 0.93, recall of 0.92, and an F1-score of 0.93. The standalone Point-of-Gaze biometric classifier achieved 82.0% accuracy. Crucially, fusing both tiers into the dual-tier ensemble yielded an overall accuracy of **94.6%**, significantly outperforming standalone visual baselines.

TABLE IV. DETECTION PERFORMANCE COMPARISON

Model / Architecture	Accuracy	Precision	Recall	F1-Score
Single-Frame Baseline (ResNet18)	84.1%	0.82	0.86	0.84
PhaseForensics [2]	86.9%	0.85	0.88	0.86
Wavelet-Conv3D Framework [3]	88.5%	0.89	0.87	0.88
PoG Gaze Biometric Alone [1]	82.0%	0.81	0.83	0.82
Proposed ResNet18Video (Visual)	**92.8%**	**0.93**	**0.92**	**0.93**
Proposed Dual-Tier Ensemble	**94.6%**	**0.95**	**0.94**	**0.94**

C. Ablation Study

To quantify the contribution of each design component, ablation experiments were conducted by systematically removing individual modules, as reported in Table V. Removing the 8-frame temporal average pooling caused the largest drop in accuracy (-8.7%), confirming that temporal aggregation is essential to suppress single-frame jitter. Eliminating 1.5× context expansion reduced accuracy by 4.2%, verifying that boundary seam information is critical for distinguishing modern face swaps.

TABLE V. ABLATION STUDY OF SYSTEM COMPONENTS

Configuration	Accuracy	F1-Score	FPR (%)
Full Dual-Tier System	**94.6%**	**0.94**	**1.8%**
Without PoG Gaze Biometrics	92.8%	0.93	2.4%
Without 8-Frame Pooling (Single Frame)	85.9%	0.85	5.8%
Without 1.5× Context Crop (Landmarks only)	90.4%	0.90	3.6%
Without Secondary Face Blurring	88.1%	0.87	6.2%
Without EMA Smoothing & Hysteresis	91.5%	0.91	4.1%

D. Latency and Execution Profiling

A central design requirement is sustaining sub-35 ms latency (≥30 FPS) for live call inspection. Table VI breaks down the execution budget across all sequential pipeline stages.

TABLE VI. LATENCY BREAKDOWN PER PROCESSING STAGE

Processing Stage	Component / Module	Latency (ms)	Budget %
Screen Ingestion	MSS Memory-Mapped Capture	6.2 ms	21.5%
Face Detection	YuNet ONNX (320×320 dynamic)	8.4 ms	29.1%
Preprocessing	Context Crop & Normalization	1.9 ms	6.6%
CNN Inference	ResNet18 Batch Forward Pass	10.6 ms	36.7%
Overlay & HUD	EMA Hysteresis & Accent Boxes	1.8 ms	6.1%
Total End-to-End	**Complete Forensic Loop**	**28.9 ms**	**100.0%**

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

With total loop execution clocking at 28.9 ms (34.6 FPS), DeepGaze-Guard comfortably satisfies interactive real-time video requirements.On CPU-only execution, average latency is 62.4 ms (16.0 FPS), remaining fully responsive for periodic background monitoring.

VII. DISCUSSION AND OPERATIONAL CONSIDERATIONS

A. Robustness to Extreme Compression and Streaming Artifacts

Video conferencing applications dynamically adjust video bitrate in response to network congestion, frequently reducing resolution from 1080p to 480p or 360p while introducing blocking artifacts. The 1.5× contextual expansion ensures that macro-level head contours and jawline discontinuities remain observable even when fine facial textures are blurred. Furthermore, Point-of-Gaze tracking operates on eye geometry rather than high-frequency texture, preserving forensic visibility under severe bandwidth throttling.

B. Prevention of Recursive Video Capture (Hall-of-Mirrors)

When forensic software renders an on-screen preview window of the captured desktop region, a hazardous recursive feedback loop can occur if the preview overlaps the source window: the detector analyzes its own UI overlays in an infinite cascade. DeepGaze-Guard neutralizes this failure mode by: (i) pre-positioning the preview window at an offset coordinate (30, 40) with fixed 480×700 bounds; and (ii) programmatically excluding the window title `Real-Time Face-Swap Detection (Output)` from the target window enumeration list.

C. Privacy, Regulation, and Ethical Considerations

Video telemetry contains sensitive personal data. DeepGaze-Guard strictly operates on volatile local memory without transmitting raw video frames or gaze logs to external cloud infrastructure, satisfying the data minimization mandates of GDPR (Article 25) and biometric compliance requirements. In compliance with the EU Artificial Intelligence Act for high-risk verification systems, the platform provides intrinsic explainability through transparent HUD confidence metrics and ocular fixation traces rather than opaque black-box classifications.

VIII .CONCLUSION AND FUTURE WORK

This paper presented **DeepGaze-Guard**, an end-to-end, explainable forensic defense platform for real-time face-swap detection during live video communications. By synthesizing high-speed spatiotemporal convolutional modeling with Point-of-Gaze behavioral biometrics, the proposed architecture overcomes the critical vulnerabilities of conventional detectors that depend exclusively on static pixel artifacts. The inclusion of ONNX-accelerated YuNet face localization with 1.5× contextual expansion, 8-frame spatiotemporal pooling, EMA confidence stabilization, and WhatsApp video pane auto-framing provides a practical, non-invasive defense directly applicable to modern communication channels. Quantitative evaluations confirm 92.8% visual accuracy, 82.0% PoG biometric accuracy, and 94.6% joint multimodal accuracy, executing with an end-to-end latency of only 28.9 ms (34.6 FPS).

Future work will advance along three trajectories: (i) integrating multimodal acoustic-phoneme to lip-motion synchronization models to detect audio dubbing; (ii) porting the convolutional backbone to TensorRT and mobile NPU runtimes for zero-overhead smartphone conferencing protection; and (iii) extending conversational turn graphs to multi-party enterprise conferences with arbitrary participant topologies.

ACKNOWLEDGMENT

The authors express their deepest gratitude to the Department of Information Science and Engineering, GM Institute of Technology, Davangere, for providing the laboratory facilities, computing infrastructure, and technical guidance necessary to complete this project. The authors also sincerely thank their project guide, faculty coordinators, and colleagues for their invaluable insights, continuous support, and constructive feedback throughout the course of this research.

REFERENCES

[1] O. M. T. Kohler, R. Vijaykumar, and M. H. Imtiaz, "DeepFake Detection in Dyadic Video Calls Using Point of Gaze Tracking," *IEEE Access*, vol. 14, pp. 11328999, 2026.

[2] E. Prashnani, M. Goebel, and B. S. Manjunath, "Generalizable Deepfake Detection with Phase-Based Motion Analysis," *IEEE Transactions on Image Processing*, vol. 34, 2025.

[3] L. A. Motalib, O. Mustapha, and H. Mustapha, "Compression-Aware Hybrid Framework for Deep Fake Detection in Low-Quality Video," *IEEE Access*, vol. 13, 2025.

[4] D. Zhu, Y. Li, Z. Wang, B. Wu, J. Zhou, and S. Lyu, "Hiding Faces in Plain Sight: Defending DeepFakes by Disrupting Face Detection," *IEEE Transactions on Dependable and Secure Computing*, 2025.

[5] A. Luo, C. Kong, X. Kang, J. Huang, Y. Hu, and A. C. Kot, "Beyond the Prior Forgery Knowledge: Mining Critical Clues for General Face Forgery Detection," *IEEE Transactions on Information Forensics and Security*, vol. 19, 2024.

[6] M. Fang, L. Yu, Y. Song, Y. Zhang, and H. Xie, "IEIRNet: Inconsistency Exploiting Based Identity Rectification for Face Forgery Detection," *IEEE Transactions on Multimedia*, vol. 26, 2024.

[7] F. Felouat et al., "eKYC-DF: A Large-Scale Dataset for Deepfake Detection in Identity Verification," *IEEE Access*, vol. 12, 2024.

[8] Y. Yu et al., "Diff-ID: An Explainable Identity Difference Quantification Framework for DeepFake Detection," *IEEE Transactions on Dependable and Secure Computing*, 2024.

[9] S. Park, H. Na, and D. Choi, "Verifiable Facial De-Identification in Video Surveillance," *IEEE Access*, vol. 12, 2024.

[10] Z. Zhao, X. Zhang, and Y. Demiris, "3PFS: Protecting Pedestrian Privacy Through Face Swapping," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, 2024.

[11] L. Zhang, T. Qiao, M. Xu, N. Zheng, and S. Xie, "Unsupervised Learning-Based Framework for Deepfake Video Detection," *IEEE Transactions on Multimedia*, vol. 25, 2023.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

[12] Y. Akbari et al., "A New Dataset for Forged Smartphone Videos Detection: Description and Analysis," *IEEE Access*, vol. 11, 2023

[13] G. Pang, B. Zhang, Z. Teng, Z. Qi, and J. Fan, "MRE-Net: Multi-Rate Excitation Network for Deepfake Video Detection," *IEEE Transactionson Circuits and Systems for Video Technology*, vol. 33, 2023

[14] S. Waseem et al., "Deep Fake on Face and Expression Swap: A Review,"*IEEE Access*, vol. 11, 2023.

[15] L. Tian et al., "Fake PoI: A Large-Scale Fake Person of Interest Video Detection Benchmark and a Strong Baseline," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[16] Y. Wang et al., "Forgery NIR: Deep Face Forgery and Detection in Near-Infrared Scenario," *IEEE Transactions on Information Forensics andSecurity*, vol. 17, 2022.

[17] P. Yu, J. Fei, Z. Xia, Z. Zhou, and J. Weng, "Improving Generalization by Commonality Learning in Face Forgery Detection," *IEEE Transactions on Information Forensics and Security*, vol. 17, 2022.

[18] X. Liu, Y. Liu, J. Chen, and X. Liu, "PSCC-Net: Progressive Spatio-Channel Correlation Network for Image Manipulation Detection andLocalization," *IEEE Transactions on Circuits and Systems for VideoTechnology*, vol. 32, 2022.

[19] L. Zhang et al., "AP-GAN: Improving Attribute Preservation in Video Face Swapping," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, 2022.

[20] A. Groshev, A. Maltseva, D. Chesakov, A. Kuznetsov, and D. Dimitrov,"GHOST: A New Face Swap Approach for Image and Video Domains,"*IEEE Access*, vol. 10, 2022.

[21] E. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "Deep Fake Detection Based on Discrepancies Between Faces and Their Context," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[22] G. Jia et al., "Inconsistency-Aware Wavelet Dual-Branch Network for Face Forgery Detection," *IEEE Transactions on Biometrics, Behavior,and Identity Science*, vol. 3, 2021

[23] M. A. Hoque, M. S. Ferdous, M. Khan, and S. Tarkoma, "Real, Forgedor Deep Fake? Enabling the Ground Truth on the Internet," *IEEE Access*,vol. 9, 2021.

[24] A. Jamsheed and J. B., "Deep Fake Video Detection Using CNN and LSTM Networks," *International Journal of Scientific Research in Computer Science and Engineering*, vol. 9, 2021

[25] M. Rafique et al., "Deepfake Detection Using Error Level Analysis and Deep Learning Techniques," *IEEE International Conference on Artificial Intelligence*, 2021.

[26] T. Baltrušaitis, P. Robinson, and L. P. Morency, "OpenFace: An opensource facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl.Comput. Vis. (WACV)*, 2016, pp. 1–10.

[27] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation for Extreme HeadPose and Gaze Variation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 162–175, 2019.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, 2016, pp. 770–778.

[29] European Commission, "Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act),"COM(2021) 206 final, 2021.

[30] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?':Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int.Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)